Target Case Study

Data Analysis using SQL

Uppinder Chugh

03.06.2024

DSML Apr 24 w/o Python

INTRODUCTION

In this data analysis report, we delve into Target's operations in Brazil by examining a dataset of 100,000 orders placed between 2016 and 2018. Through this analysis, we aim to uncover key insights into order processing, pricing strategies, payment and shipping efficiency, customer demographics, product characteristics, and customer satisfaction levels.

SOLUTION

1. **Import the dataset and do usual exploratory analysis steps like checking the structure & characteristics of the dataset.**
   A. Data type of all columns in the “customers” table.

```
SELECT
 column_name,
 data_type
FROM
 `target.INFORMATION_SCHEMA.COLUMNS`
WHERE
 TABLE_NAME = "customers"
```

Output:

Row	column_name	data_type
1	customer_id	STRING
2	customer_unique_id	STRING
3	customer_zip_code_prefix	INT64
4	customer_city	STRING
5	customer_state	STRING

   B. Get the time range between which the orders were placed.

```
SELECT
  *,
  DATE_DIFF(last_transaction,first_transaction,day) AS order_period
FROM (
  SELECT
    MIN(order_purchase_timestamp) AS first_transaction,
    MAX(order_purchase_timestamp) AS last_transaction
```

```
  FROM
    `target.orders`
) AS subquery;
```

Output:

Row	first_transaction	last_transaction	order_period
1	2016-09-04 21:15:19 UTC	2018-10-17 17:30:18 UTC	772

Observation: The orders were placed from **4 Sept 2016 to 17 Nov 2018** which span around **772** days.

C. Count the Cities & States of customers who ordered during the given period.

```
SELECT
  count(distinct customer_state) as num_of_states,
  count(distinct customer_city) as num_of_cities
FROM
  `target.orders` inner join `target.customers` using(customer_id);
```

Output:

Row	num_of_states	num_of_cities
1	27	4119

The number of states and cities helps us understand the business's area of coverage. The number of states and cities a retail business operates in reveals its geographic reach and market penetration, allowing it to serve a diverse customer base and potentially increase revenue.

2. **In-depth Exploration:**
   A. Is there a growing trend in the no. of orders placed over the past years?

```
SELECT
  FORMAT_TIMESTAMP("%Y-%m", order_purchase_timestamp) AS year_month,
  COUNT(*) AS num_of_orders
FROM
  `target.orders`
GROUP BY
  year_month
ORDER BY
  year_month;
```

Output:

Row	year_month	num_of_orders
1	2016-09	4
2	2016-10	324
3	2016-12	1
4	2017-01	800
5	2017-02	1780
6	2017-03	2682
7	2017-04	2404
8	2017-05	3700
9	2017-06	3245
10	2017-07	4026

Chart:

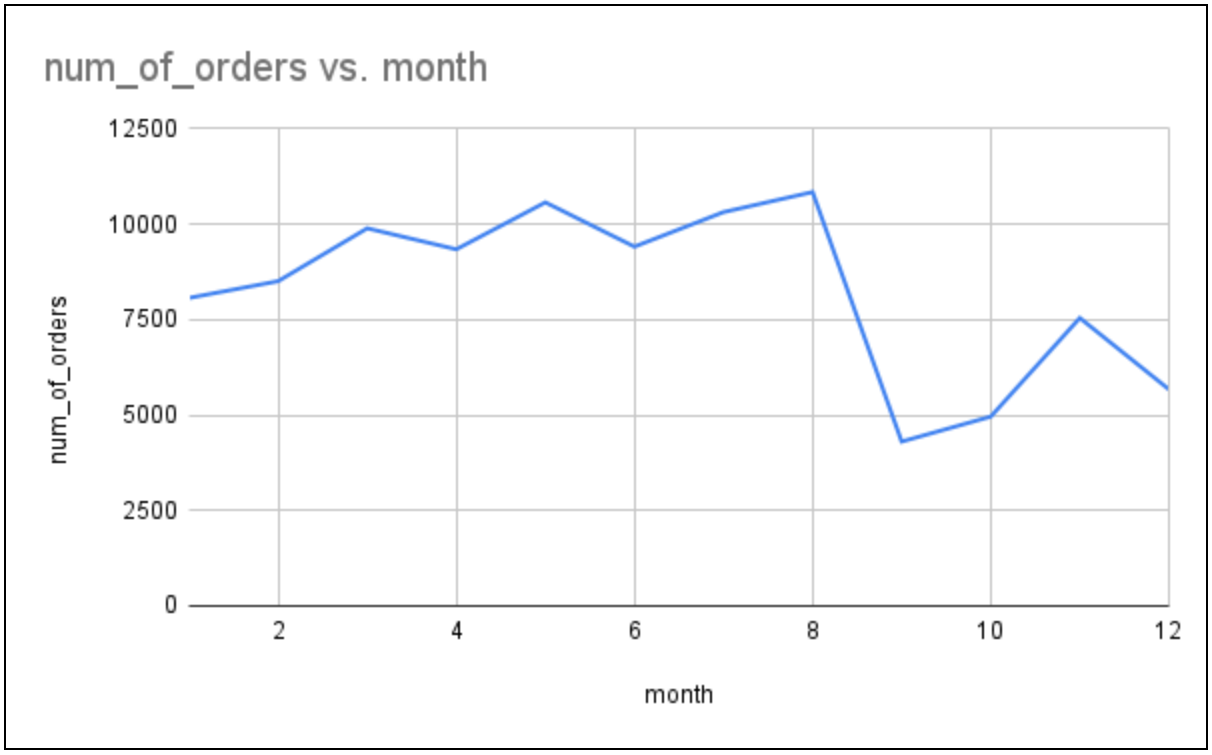
Observation: The increasing trend in the number of orders over the past years signifies growing customer demand and market acceptance, which is crucial for business growth and sustainability.

B. Can we see some kind of monthly seasonality in terms of the no. of orders being placed?

```
SELECT
  EXTRACT(month FROM order_purchase_timestamp) AS month,
  COUNT(DISTINCT order_id) AS num_of_orders
FROM `target.orders`
GROUP BY month
ORDER BY month;
```

Output:

Row	month	num_of_orders
1	1	8069
2	2	8508
3	3	9893
4	4	9343
5	5	10573
6	6	9412
7	7	10318
8	8	10843
9	9	4305
10	10	4959



Observation: The line chart shows monthly seasonality in the number of orders. Order volume gradually increases from January to a peak in March-April, fluctuates through the middle months, drops sharply in August, recovers in September-October, and decreases again towards December.

C. During what time of the day, do the Brazilian customers mostly place their orders? (Dawn, Morning, Afternoon or Night) period.

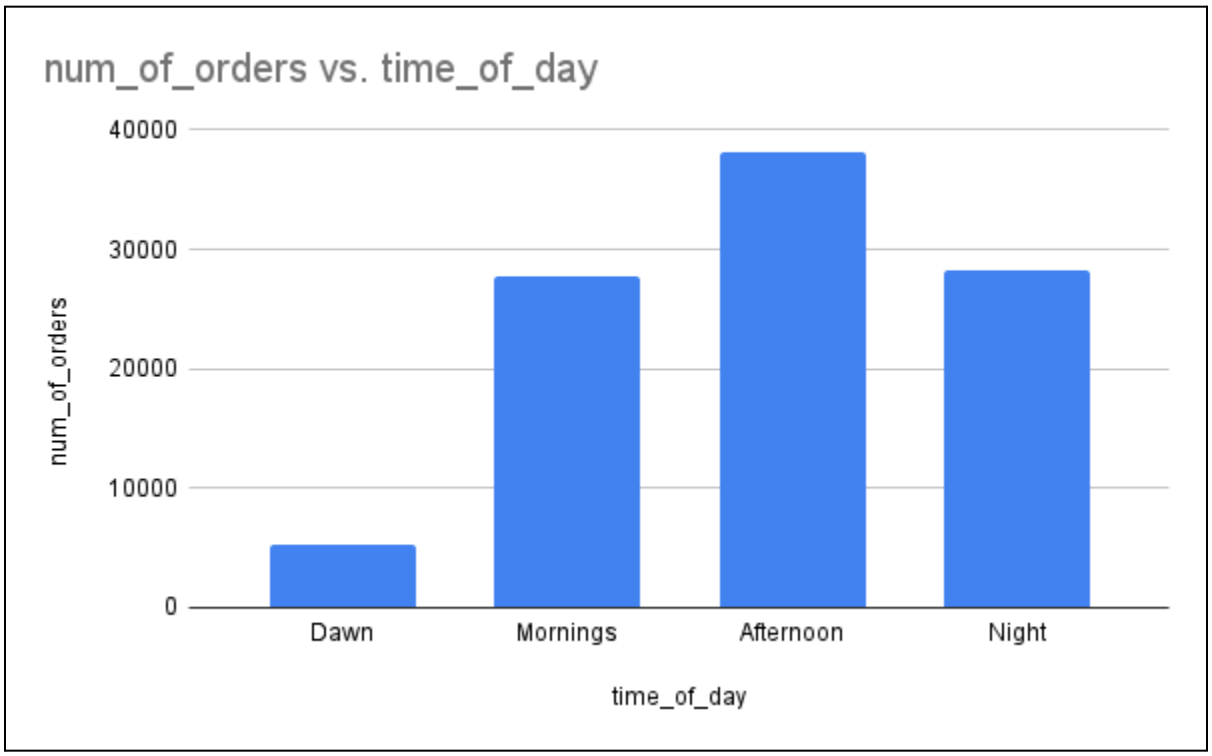
```
WITH hour_cte AS (
  SELECT
      EXTRACT(hour FROM order_purchase_timestamp) AS order_hour
    FROM
      `target.orders`
)

SELECT
  (
    CASE
      WHEN order_hour BETWEEN 0 AND 6 THEN "Dawn"
      WHEN order_hour BETWEEN 7 AND 12 THEN "Mornings"
      WHEN order_hour BETWEEN 13 AND 18 THEN "Afternoon"
      ELSE "Night"
  END
    ) AS time_of_day,
  COUNT(*) AS num_of_orders
FROM hour_cte
GROUP BY
  time_of_day
ORDER BY
  time_of_day;
```

Output:

Row	time_of_day	num_of_orders
1	Afternoon	38135
2	Dawn	5242
3	Mornings	27733
4	Night	28331

Chart:



Observation: We observe a pattern where Brazilian customers place fewer orders during dawn, the highest number during the afternoon, and nearly equivalent amounts during mornings and nights. Understanding the customer ordering patterns allows the business to optimize staffing and resources, ensuring sufficient support during peak hours while minimizing costs during slower periods.

3. **Evolution of E-commerce orders in the Brazil region:**
   A. Get the month-on-month no. of orders placed in each state.

```
SELECT
  customer_state,
  EXTRACT(month FROM order_purchase_timestamp) AS order_month,
  COUNT(*) AS num_of_orders
FROM `target.orders` JOIN `target.customers` USING(customer_id)
GROUP BY customer_state,order_month
ORDER BY customer_state,order_month;
```

Output:

Row	customer_state	order_month	num_of_orders
3	AC	3	4
4	AC	4	9
5	AC	5	10
6	AC	6	7
7	AC	7	9
8	AC	8	7
9	AC	9	5
10	AC	10	6
11	AC	11	5
12	AC	12	5

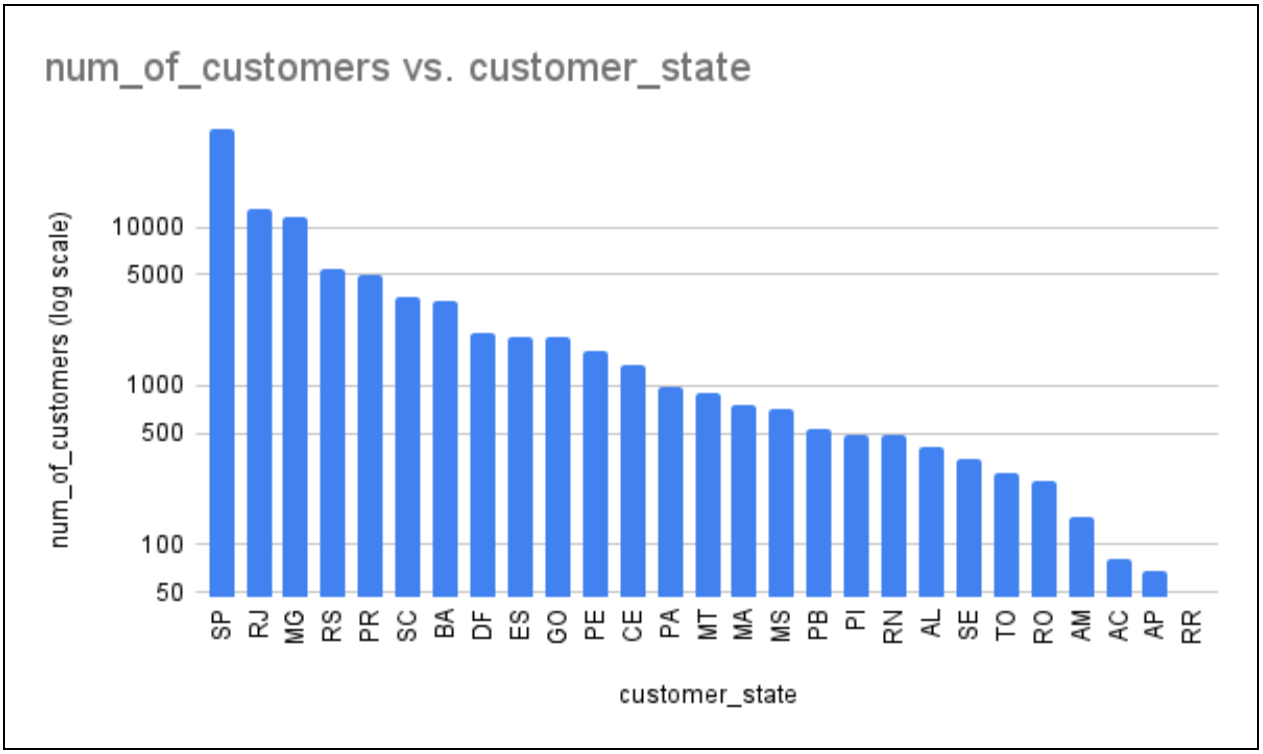
   B. How are the customers distributed across all the states?

```
SELECT
  customer_state,
  COUNT(customer_unique_id) AS num_of_customers
FROM
  `target.customers`
GROUP BY
  customer_state
ORDER BY
  num_of_customers DESC;
```

Output:

Row	customer_state	num_of_customers
1	SP	41746
2	RJ	12852
3	MG	11635
4	RS	5466
5	PR	5045
6	SC	3637
7	BA	3380
8	DF	2140
9	ES	2033
10	GO	2020

Chart:



Observation: We can observe an exponential distribution in the number of customers by state, with São Paulo (SP) having the highest count at 41,746 and Roraima (RR) the lowest at 46.

Understanding how customers are spread out across states helps the business know which areas have more people buying their products. This helps them decide where to focus their advertising and resources to make the most money, especially in big places like São Paulo. They can also make sure they're reaching customers in smaller places like Roraima.

4. **Impact on Economy: Analyze the money movement by e-commerce by looking at order prices, freight and others.**
    A. Get the % increase in the cost of orders from year 2017 to 2018 (include months between Jan to Aug only).

```
WITH
  order_costs AS (
    SELECT order_id,order_purchase_timestamp,SUM(payment_value) AS order_cost
    FROM `target.orders`JOIN `target.payments` USING(order_id)
    WHERE order_status != "canceled"
    GROUP BY 1,2
  ),
  yearly_totals AS (
    SELECT
      SUM(CASE WHEN order_purchase_timestamp BETWEEN '2017-01-01' AND '2017-08-31' THEN
order_cost ELSE 0 END) AS total_cost_2017,
      SUM(CASE WHEN order_purchase_timestamp BETWEEN '2018-01-01' AND '2018-08-31' THEN
order_cost ELSE 0 END) AS total_cost_2018,
    FROM order_costs
  )

SELECT
  FORMAT("%'.2f", total_cost_2017) as total_cost_2017,
  FORMAT("%'.2f", total_cost_2018) as total_cost_2018,
  CONCAT(ROUND(100.0 * (total_cost_2018 - total_cost_2017)/total_cost_2017,2),"%") AS
percentage_increase
FROM yearly_totals
```

Note: Canceled orders have been excluded from the calculation.

Output:

w	total_cost_2017	total_cost_2018	percentage_increase
1	3,604,146.58	8,624,464.94	139.29%

The significant increase of **139.29%** in the cost of orders from 2017 to 2018 between January and August indicates a substantial rise in expenses during that period, potentially impacting the business's profitability and financial planning. Understanding this insight helps the business adapt its budgeting and pricing strategies accordingly, ensuring it remains financially stable and competitive in the market.

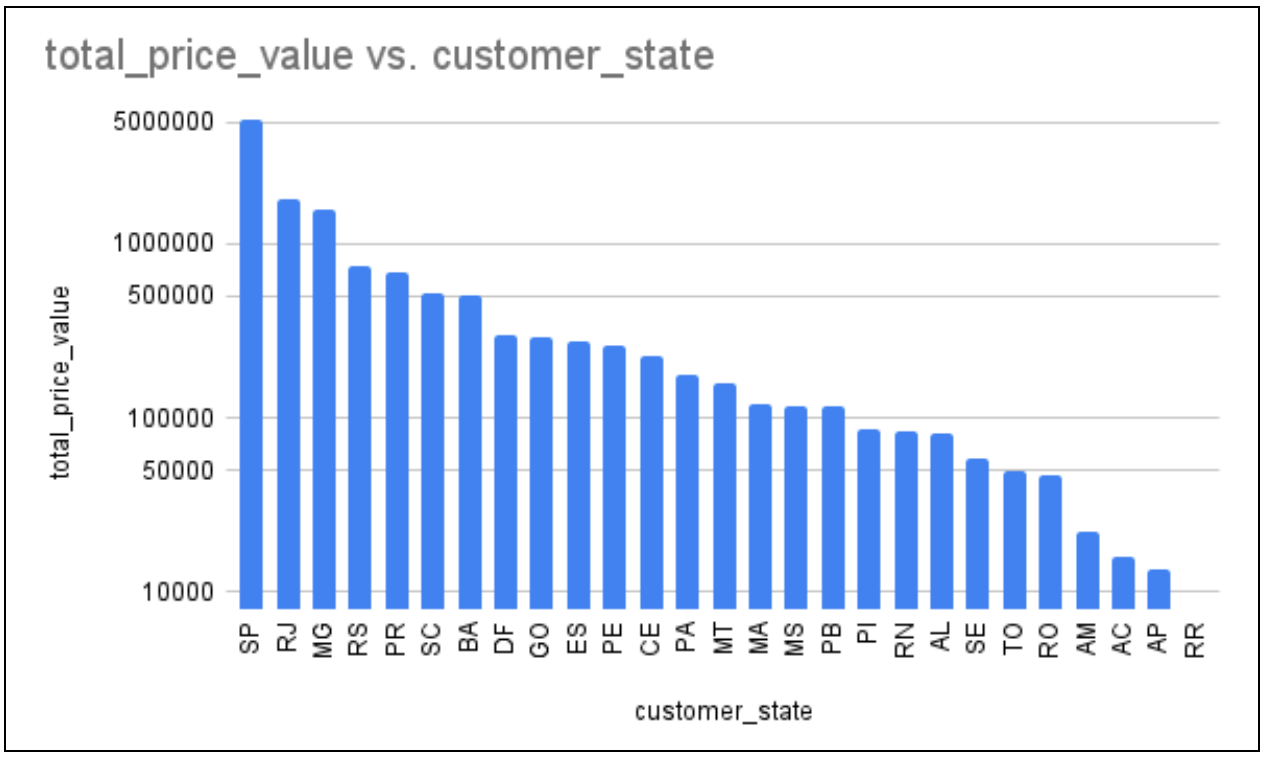
B. Calculate the Total & Average value of order price for each state.

```
SELECT
  customer_state,
  ROUND(SUM(price),2) AS total_price_value,
  ROUND(AVG(price),2) AS avg_price_value
FROM
  `target.orders`
  JOIN `target.order_items` USING(order_id)
  JOIN `target.customers` USING (customer_id)
GROUP BY customer_state
ORDER BY total_price_value DESC;
```

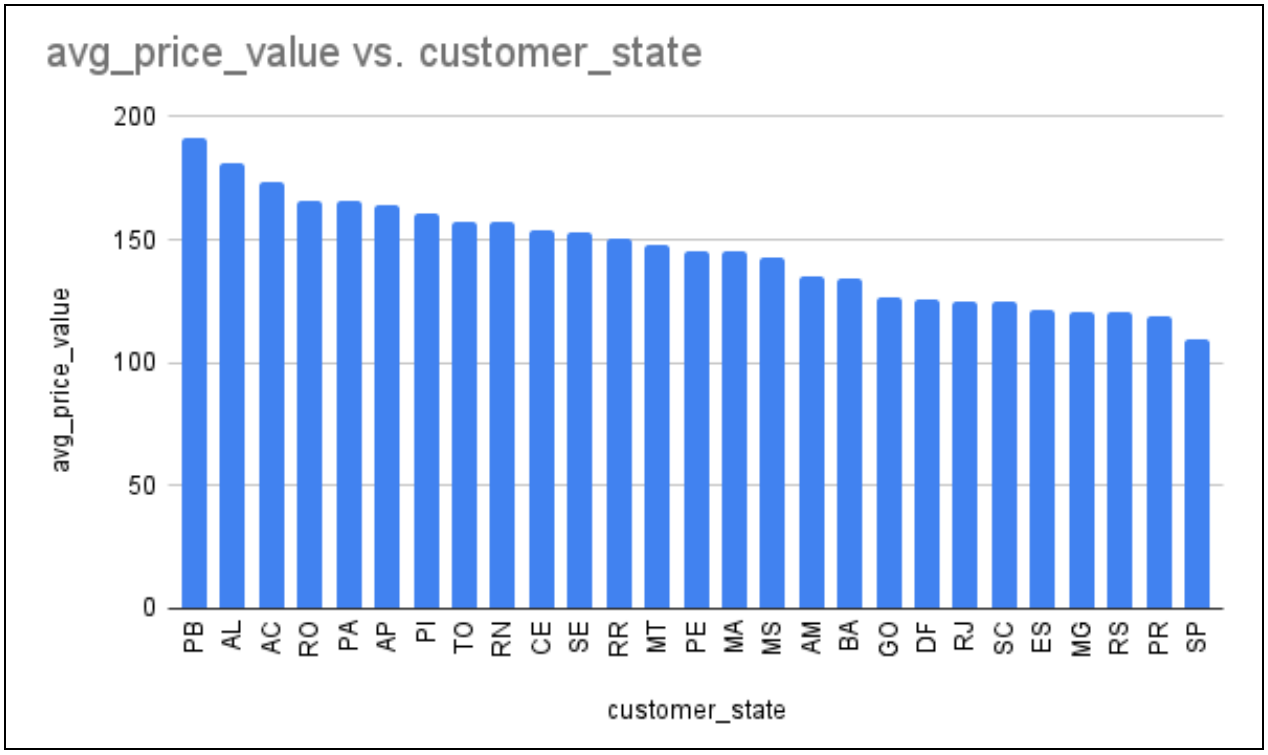
Output:

Row	customer_state	total_price_value	avg_price_value
1	SP	5202955.05	109.65
2	RJ	1824092.67	125.12
3	MG	1585308.03	120.75
4	RS	750304.02	120.34
5	PR	683083.76	119.0
6	SC	520553.34	124.65
7	BA	511349.99	134.6
8	DF	302603.94	125.77
9	GO	294591.95	126.27
10	ES	275037.31	121.91

Chart:



Observation: We can observe that the total price value by state follows an exponential distribution, with São Paulo (SP) having the highest value at 5,202,955.05 and Roraima (RR) the lowest at 7,829.43. This insight helps the business focus its efforts on high-value areas like São Paulo to maximize profits while still addressing the needs of smaller markets like Roraima.



Observation: In terms of the average payment value, there is a linear relationship by state. Paraíba (PB) has the highest average at 191.48, while São Paulo (SP) has an average of 109.65. This insight helps the business tailor its pricing and marketing strategies to different regions, focusing on maximizing sales where customers are willing to pay more.

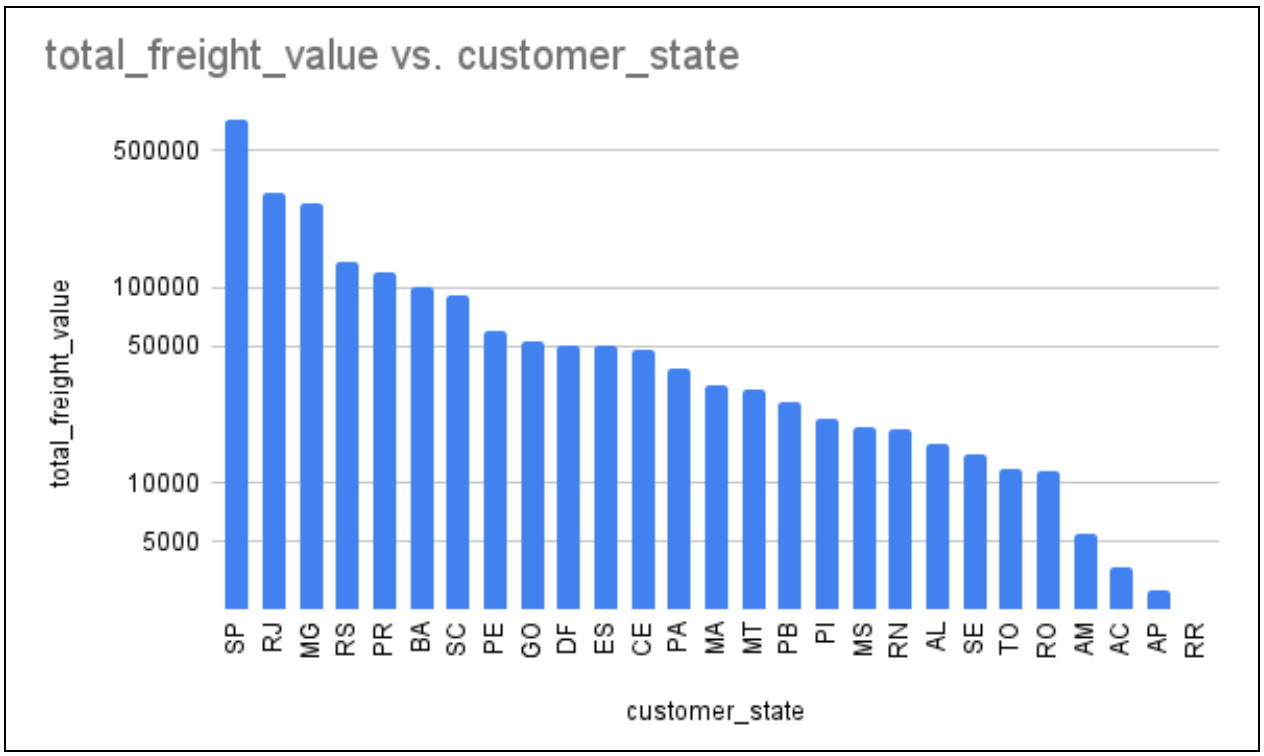
C. Calculate the Total & Average value of order freight for each state.

```
SELECT
  customer_state,
  ROUND(SUM(freight_value),1) AS total_freight_value,
  ROUND(AVG(freight_value),1) AS avg_freight_value
FROM
  `target.orders`
  JOIN `target.order_items` USING(order_id)
  JOIN `target.customers` USING (customer_id)
GROUP BY
  customer_state
ORDER BY
  total_freight_value DESC;
```

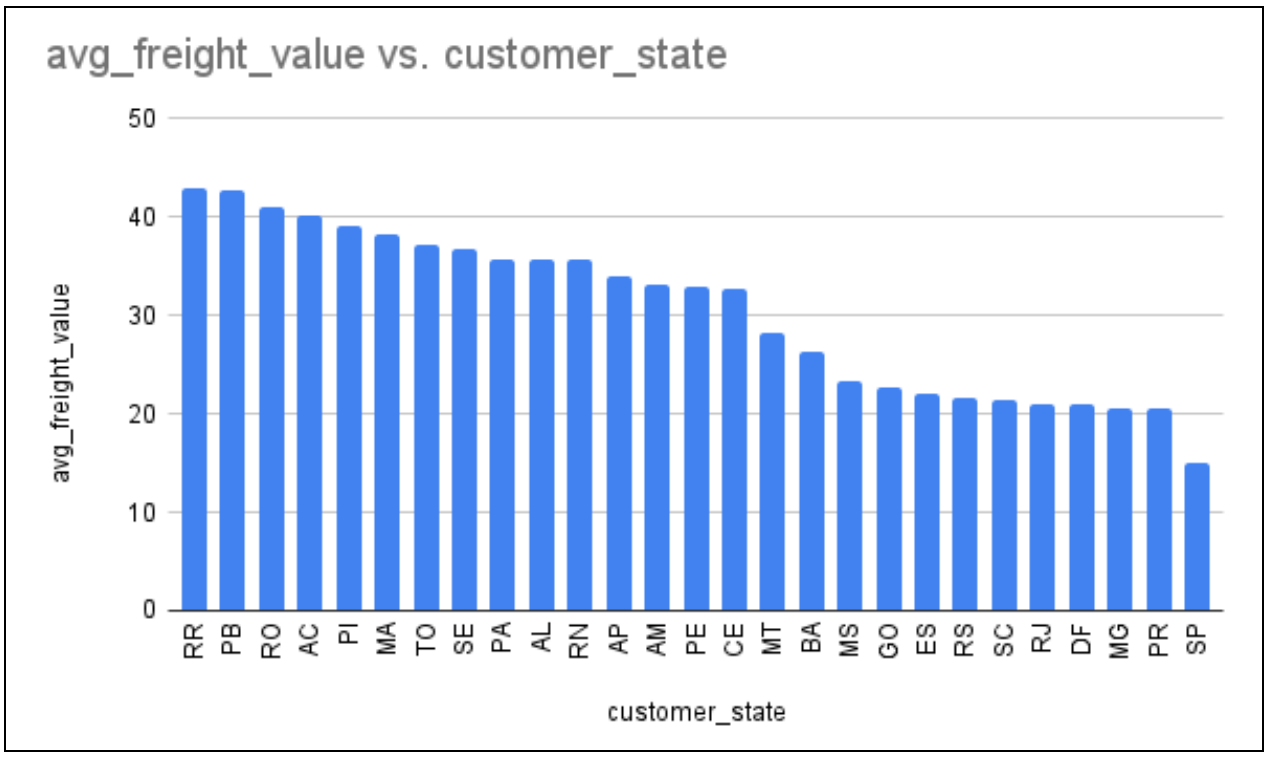
Output:

Row	customer_state	total_freight_value	avg_freight_value
1	SP	718723.1	15.1
2	RJ	305589.3	21.0
3	MG	270853.5	20.6
4	RS	135522.7	21.7
5	PR	117851.7	20.5
6	BA	100156.7	26.4
7	SC	89660.3	21.5
8	PE	59449.7	32.9
9	GO	53115.0	22.8
10	DF	50625.5	21.0

Chart:



Observation: Similar to payment values, we can observe that the total freight value by state follows an exponential distribution, with São Paulo (SP) having the highest value at 718,723.1 and Roraima (RR) the lowest at 2,235.2. This indicates that shipping costs are significantly higher in more populous areas. This insight helps the business manage its logistics more effectively by efficiently allocating resources and planning for higher shipping expenses in states with more orders.



Observation: In terms of average freight value, there is a linear relationship by state. Roraima (RR) has the highest average at 43.0, while São Paulo (SP) has an average of 15.1. Roraima has the highest average freight cost per order, while São Paulo has a lower average. This insight helps the business understand that shipping to less populated areas like Roraima is more expensive per order, which can inform pricing and shipping policies.

5. **Analysis based on sales, freight and delivery time.**
   A. Find the no. of days taken to deliver each order from the order's purchase date as delivery time. Also, calculate the difference (in days) between the estimated & actual delivery date of an order.
      - **time_to_deliver** = order_delivered_customer_date - order_purchase_timestamp
      - **diff_estimated_delivery** = order_estimated_delivery_date - order_delivered_customer_date

```
SELECT
  order_id,
  DATE_DIFF(order_delivered_customer_date, order_purchase_timestamp, day) AS
time_to_deliver,
  DATE_DIFF(order_estimated_delivery_date, order_delivered_customer_date, day) AS
diff_estimated_delivery
FROM
  `target.orders`
WHERE
  order_status = 'delivered';
```

Output:

Row	order_id	time_to_deliver	diff_estimated_delivery
1	635c894d068ac37e6e03dc54e...	30	1
2	3b97562c3aee8bdedcb5c2e45...	32	0
3	68f47f50f04c4cb6774570cfde...	29	1
4	276e9ec344d3bf029ff83a161c...	43	-4
5	54e1a3c2b97fb0809da548a59...	40	-4
6	fd04fa4105ee8045f6a0139ca5...	37	-1
7	302bb8109d097a9fc6e9cefc5...	33	-5
8	66057d37308e787052a32828...	38	-6
9	19135c945c554eebfd7576c73...	36	-2
10	4493e45e7ca1084efcd38ddeb...	34	0

Note: Negative *diff_estimated_delivery* is for orders that were delivered *after* the estimated delivery date.

B. Find out the top 5 states with the highest & lowest average freight value.

```
WITH state_freight_value_cte AS (
  SELECT
    customer_state,
    ROUND(AVG(freight_value),1) AS avg_freight_value
  FROM
    `target.orders`
    JOIN `target.order_items` USING(order_id)
    JOIN `target.customers` USING (customer_id)
  GROUP BY
    customer_state
)

SELECT customer_state,avg_freight_value
FROM state_freight_value_cte
ORDER BY avg_freight_value DESC
LIMIT 5

SELECT customer_state,avg_freight_value
FROM state_freight_value_cte
ORDER BY avg_freight_value ASC
LIMIT 5
```

Output:

Row	customer_state	avg_freight_value
1	RR	43.0
2	PB	42.7
3	RO	41.1
4	AC	40.1
5	PI	39.1

Top 5 states with the highest average freight value

Row	customer_state	avg_freight_value
1	SP	15.1
2	PR	20.5
3	MG	20.6
4	RJ	21.0
5	DF	21.0

Top 5 states with the lowest average freight value

C. Find out the top 5 states with the highest & lowest average delivery time.

```
WITH statewise_avg_delivery_time_cte AS (
  SELECT
    customer_state,
    AVG(DATE_DIFF(order_delivered_customer_date, order_purchase_timestamp, day)) AS
avg_delivery_time,
  FROM
  `target.orders` JOIN `target.customers` USING(customer_id)
  WHERE order_status = 'delivered'
  GROUP BY customer_state
)

-- Highest
SELECT customer_state,ROUND(avg_delivery_time,2) as avg_delivery_time
FROM statewise_avg_delivery_time_cte
ORDER BY avg_delivery_time DESC
LIMIT 5

-- Lowest
SELECT customer_state,ROUND(avg_delivery_time,2) as avg_delivery_time
FROM statewise_avg_delivery_time_cte
ORDER BY avg_delivery_time ASC
LIMIT 5
```

Output:

Row	customer_state	avg_delivery_time
1	RR	28.98
2	AP	26.73
3	AM	25.99
4	AL	24.04
5	PA	23.32

Top 5 states with the highest average delivery time

Row	customer_state	avg_delivery_time
1	SP	8.3
2	PR	11.53
3	MG	11.54
4	DF	12.51
5	SC	14.48

Top 5 states with the lowest average delivery time

Observation: As we can observe, Roraima (RR) has the highest average delivery time at **28.98** days, whereas São Paulo (SP) has the lowest average delivery time at **8.3** days. This insight highlights the need for the business to improve delivery efficiency in remote areas like Roraima and maintain quick delivery times in major areas like São Paulo.

D. Find out the top 5 states where the order delivery is really fast as compared to the estimated date of delivery.

```
SELECT
  customer_state,
  ROUND(AVG(DATE_DIFF(
            order_estimated_delivery_date,
            order_delivered_customer_date, day)),2) AS avg_diff_estimated_delivery
FROM `target.orders` JOIN `target.customers` USING(customer_id)
WHERE order_status = 'delivered'
GROUP BY customer_state
ORDER BY avg_diff_estimated_delivery DESC
LIMIT 5
```

Output:

Row	customer_state	avg_diff_estimated_delivery
1	AC	19.76
2	RO	19.13
3	AP	18.73
4	AM	18.61
5	RR	16.41

Top 5 states where the order delivery is really fast as compared to the estimated date of delivery.

Observation: Acre's orders are delivered much faster than estimated, with an average difference of nearly 20 days. This insight shows that the business can set more accurate delivery expectations, improving customer satisfaction and trust in those regions.

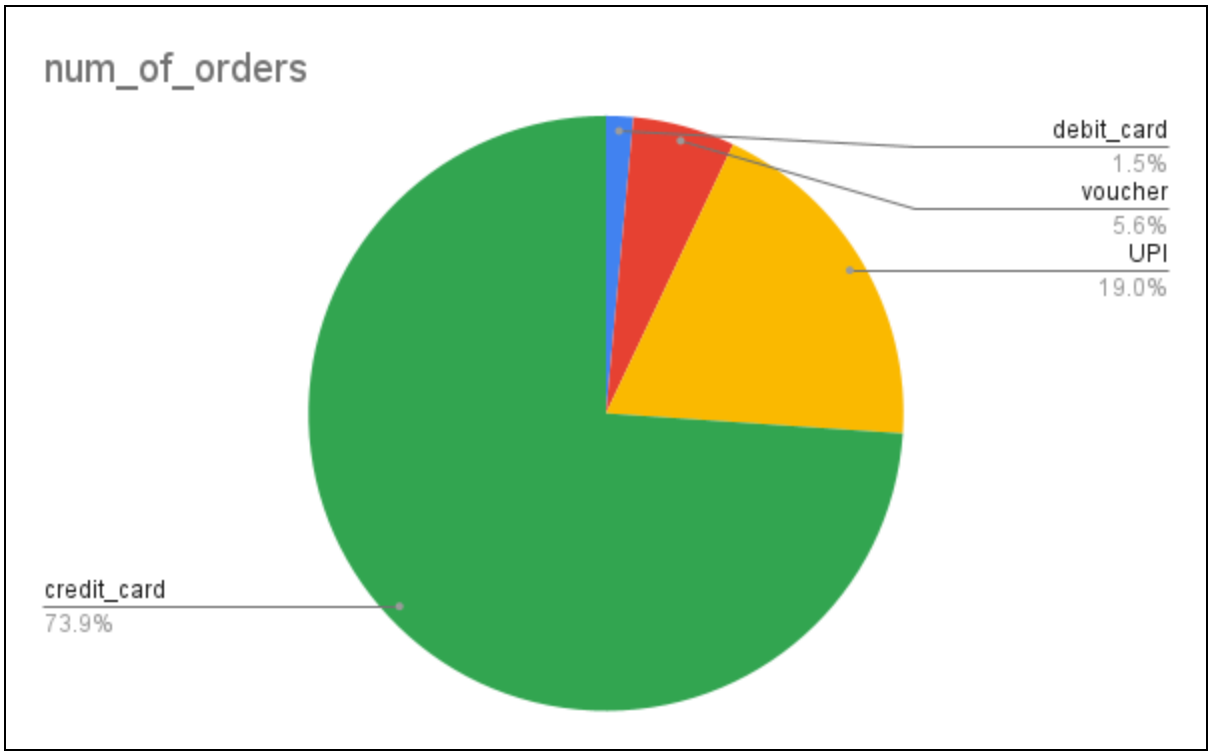
6. **Analysis based on the payments:**
    A. Find the month-on-month no. of orders placed using different payment types.

```
SELECT
  FORMAT_TIMESTAMP("%Y-%m", order_purchase_timestamp) AS year_month,
  payment_type,
  COUNT(order_id) AS num_of_orders
FROM
  `target.orders` JOIN `target.payments` USING(order_id)
GROUP BY 1,2
ORDER BY 1,2
```

Output:

Row	year_month	payment_type	num_of_orders
1	2016-09	credit_card	3
2	2016-10	UPI	63
3	2016-10	credit_card	254
4	2016-10	debit_card	2
5	2016-10	voucher	23
6	2016-12	credit_card	1
7	2017-01	UPI	197
8	2017-01	credit_card	583
9	2017-01	debit_card	9
10	2017-01	voucher	61

Studying monthly order numbers with different payment methods helps the business learn about payment trends. This helps the business adjust payment strategies and improve services to make customers happier and financial operations smoother.



Number of orders according to payment_type

The dominance of credit card payments, comprising 73.9% of all orders, suggests that customers prefer this payment method. Understanding these preferences allows the business to prioritize features and promotions related to credit card payments, potentially increasing sales and customer satisfaction.

B. Find the no. of orders placed on the basis of the payment installments that have been paid.

```
WITH
  payments_cte AS (
    SELECT DISTINCT order_id,payment_installments
    FROM `target.payments`
    WHERE payment_installments >= 1
  )

SELECT
  payment_installments,
  COUNT(DISTINCT order_id) AS num_orders
FROM payments_cte
GROUP BY 1
ORDER BY 2 DESC
```

Output:

Row	payment_installment	num_orders
1	1	49060
2	2	12389
3	3	10443
4	4	7088
5	10	5315
6	5	5234
7	8	4253
8	6	3916
9	7	1623
10	9	644

Observation: The distribution of payment installment options provides insight into customers' preferred payment plans. With a large number of orders split into 1,2 or 3 installments, it suggests that customers prefer shorter payment terms. Understanding these preferences enables the business to tailor its payment options to better suit customer needs, potentially increasing conversion rates and overall customer satisfaction.

7. **Insights and recommendations:**
    - The number of customers per state directly correlates with the total cost of orders per state. The more the number of customers in a state, the more the revenue.
    - The average freight value and average delivery time are inversely proportional to the number of customers per state.
    - According to Table 5d, the estimated delivery dates are significantly later than the actual delivery dates for some states. This discrepancy means customers expect their orders to arrive much later than they actually do, indicating that we need to be more competitive in delivery times.
    - Faster delivery times lead to higher customer reviews for delivered orders. Additionally, the greater the difference between the estimated and actual delivery dates, the higher the customer satisfaction.

```
-- Review Score in relation to the average delivery time and average difference from the
estimated delivery date
SELECT
  t.review_score,
  ROUND(AVG(DATE_DIFF(o.order_delivered_customer_date, o.order_purchase_timestamp, day)),1)
AS avg_time_to_deliver,
  ROUND(AVG(DATE_DIFF(o.order_estimated_delivery_date, o.order_delivered_customer_date,
day)),1) AS avg_diff_estimated_delivery
FROM
  target.order_reviews AS t
  JOIN target.orders AS o USING(order_id)
  JOIN target.customers AS c USING(customer_id)
WHERE o.order_delivered_customer_date IS NOT NULL
GROUP BY 1
ORDER BY 1
```

Row	review_score	avg_time_to_deliver	avg_diff_estimated_delivery
1	1	20.8	3.4
2	2	16.2	7.8
3	3	13.8	9.9
4	4	11.8	11.4
5	5	10.2	12.4

- The average delivery time for the most ordered products is more than the average time it takes for an order to reach the customer among all orders.

```
SELECT
  product_id,
  COUNT(order_id) AS num_orders,
  ROUND(AVG(DATE_DIFF(order_delivered_customer_date, order_purchase_timestamp, day)),1) AS
avg_time_to_deliver,
  (
    SELECT
      ROUND(AVG(DATE_DIFF(order_delivered_customer_date, order_purchase_timestamp, day)),1)
    FROM target.orders
    WHERE order_delivered_customer_date IS NOT NULL
  ) AS overall_avg_time_to_deliver
FROM
  target.order_items JOIN target.orders USING(order_id)
WHERE order_delivered_customer_date IS NOT NULL
GROUP BY 1
ORDER BY 2 desc
```

Row	product_id	num_orders	avg_time_to_deliver	overall_avg_time_to_deliver
1	aca2eb7d00ea1a7b8ebd4e683…	520	11.4	12.1
2	422879e10f46682990de24d77…	484	15.4	12.1
3	99a4788cb24856965c36a24e3…	477	13.1	12.1
4	389d119b48cf3043d311335e4…	390	15.3	12.1
5	368c6c730842d78016ad8238…	388	15.0	12.1
6	53759a2ecddad2bb87a079a1f…	373	15.3	12.1
7	d1c427060a0f73f6b889a5c7c…	332	12.9	12.1
8	53b36df67ebb7c41585e8d54d…	321	14.1	12.1
9	154e7e31ebfa092203795c972…	274	9.3	12.1
10	3dd2a17168ec895c781a9191c…	272	12.7	12.1

Average Delivery Time for Most Ordered Products Compared with Overall Average Delivery Time for All Products

For most of the top-ordered products, having a higher average delivery time compared to the overall average suggests potential inefficiencies or delays in handling these popular items. This insight underscores the importance for the business to address logistical bottlenecks or optimize processes to ensure timely delivery, maintaining customer satisfaction and loyalty.

- Comparison of Average Review Score by Product Category and Overall Average Review Score

```
SELECT
  product_id,
  COUNT(order_id) AS num_orders,
  ROUND(AVG(DATE_DIFF(order_delivered_customer_date, order_purchase_timestamp, day)),1) AS
avg_time_to_deliver,
  (
    SELECT
      ROUND(AVG(DATE_DIFF(order_delivered_customer_date, order_purchase_timestamp, day)),1)
    FROM target.orders
    WHERE order_delivered_customer_date IS NOT NULL
  ) AS overall_avg_time_to_deliver
FROM
  target.order_items JOIN target.orders USING(order_id)
WHERE order_delivered_customer_date IS NOT NULL
GROUP BY 1
ORDER BY 2 DESC
```

Row	product_category	num_orders	avg_review_score	overall_avg_review_score
1	bed table bath	11137	3.9	4.0
2	Furniture Decoration	8331	3.9	4.0
3	computer accessories	7849	3.9	4.0
4	telephony	4517	3.9	4.0
5	Furniture office	1687	3.5	4.0

The observation that product categories with the highest order volumes have average review scores lower than the overall average suggests potential dissatisfaction or issues with these popular items. This insight highlights the need for the company to address quality or service concerns within these categories to improve customer satisfaction and maintain market competitiveness.

- State-wise Average product review

```
SELECT
  customer_state,
  ROUND(AVG(review_score),2) AS avg_review_score
FROM
  target.order_reviews
  JOIN target.orders USING(order_id)
  JOIN target.customers USING(customer_id)
GROUP BY 1
ORDER BY 2
```

Row	customer_state	avg_review_score
1	RR	3.61
2	AL	3.75
3	MA	3.76
4	SE	3.81
5	PA	3.85
6	CE	3.85
7	BA	3.86
8	RJ	3.87
9	PI	3.92
10	PE	4.01

Roraima's lower average review score of 3.61 compared to Amapá's 4.19 suggests that customers in Roraima are less satisfied. This insight indicates a need for the business to investigate and address the factors contributing to lower satisfaction in Roraima to improve overall customer experience.

Recommendations:

1. **Streamline Popular Product Delivery:** Make sure top-selling items are delivered efficiently by fixing any delays or issues in the shipping process, ensuring customers get their orders on time.

2. **Improve Quality in High-Demand Categories:** Focus on making products in popular categories better to satisfy customers and keep them coming back for more.

3. **Be More Accurate with Delivery Dates:** Work on making estimated delivery dates more accurate so customers aren't surprised when their orders arrive earlier or later than expected.

4. **Address Customer Concerns in Low-Rated Areas:** Figure out why customers in places like Roraima are less satisfied and fix those issues to make sure everyone's happy with their purchases.

5. **Prioritize Customer Feedback:** Listen to what customers are saying, especially in regions with lower satisfaction ratings like Roraima, and take action to address their concerns, ensuring their voices are heard and their needs are met.

CONCLUSION

In conclusion, the analysis of Target's operations in Brazil reveals key insights into customer ordering patterns, regional demand, payment preferences, and delivery efficiencies. Understanding these trends helps Target optimize its logistics, marketing strategies, and customer service to better meet the needs of its diverse customer base. By leveraging these insights, Target can enhance its operational efficiency and customer satisfaction across Brazil.